GREAT LEARNING

PGP-DSBA Online

Feb'21 Batch

Date: 26/06/2021

References: Kmeans Clustering.pdf, Clustering Monograph DSBA.pdf

Submitted By: Prachi Gupta

CLUSTERING

Problem Statement:

A leading bank wants to develop a customer segmentation to give promotional offers to its customers. They collected a sample that summarizes the activities of users during the past few months. You are given the task to identify the segments based on credit card usage.

Attribute Information:

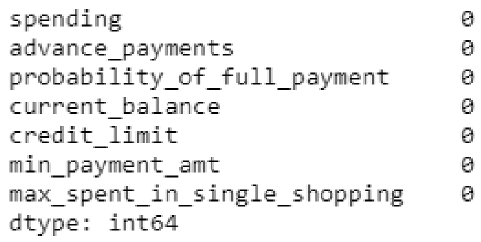
1. **spending:** Amount spent by the customer per month (in 1000s)
2. **advance_payments:** Amount paid by the customer in advance by cash (in 100s)
3. **probability_of_full_payment:** Probability of payment done in full by the customer to the bank
4. **current_balance:** Balance amount left in the account to make purchases (in 1000s)
5. **credit_limit:** Limit of the amount in credit card (10000s)
6. **min_payment_amt:** minimum paid by the customer while making payments for purchases made monthly (in 100s)
7. **max_spent_in_single_shopping:** Maximum amount spent in one purchase (in 1000s)

Problem 1:

1.1 Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).

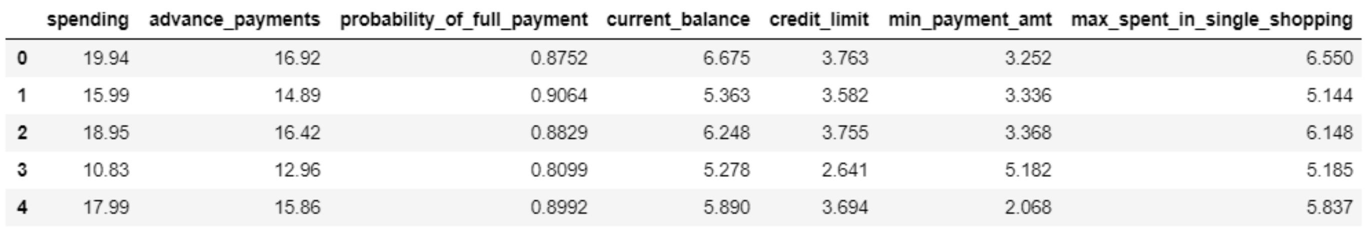
Pre-processing Data:

- This step checks for dataset size, shape,checking data types of variables present, handling missing and duplicate values.Below is a glance of the datset used (**Figure: 2)**
- There are 210 rows and 7 columns in our dataset.
- None of the columns have any null/missing value(**Figure: 1)**
- There are no duplicate values present in our dataset.



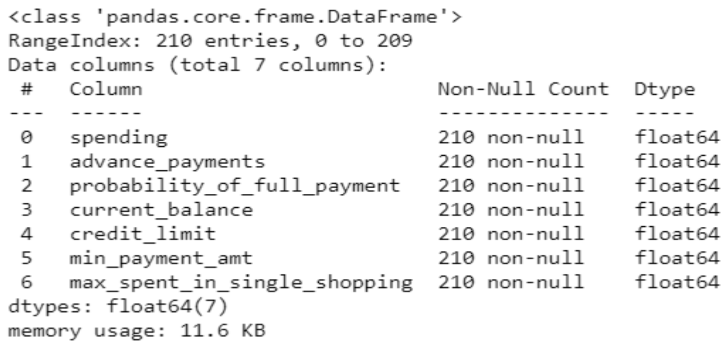
```
spending                        0
advance_payments                0
probability_of_full_payment     0
current_balance                 0
credit_limit                    0
min_payment_amt                 0
max_spent_in_single_shopping    0
dtype: int64
```

Figure: 1



	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
0	19.94	16.92	0.8752	6.675	3.763	3.252	6.550
1	15.99	14.89	0.9064	5.363	3.582	3.336	5.144
2	18.95	16.42	0.8829	6.248	3.755	3.368	6.148
3	10.83	12.96	0.8099	5.278	2.641	5.182	5.185
4	17.99	15.86	0.8992	5.890	3.694	2.068	5.837

Figure: 2



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 210 entries, 0 to 209
Data columns (total 7 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   spending                      210 non-null    float64
 1   advance_payments              210 non-null    float64
 2   probability_of_full_payment   210 non-null    float64
 3   current_balance               210 non-null    float64
 4   credit_limit                  210 non-null    float64
 5   min_payment_amt               210 non-null    float64
 6   max_spent_in_single_shopping  210 non-null    float64
dtypes: float64(7)
memory usage: 11.6 KB
```

Figure: 3

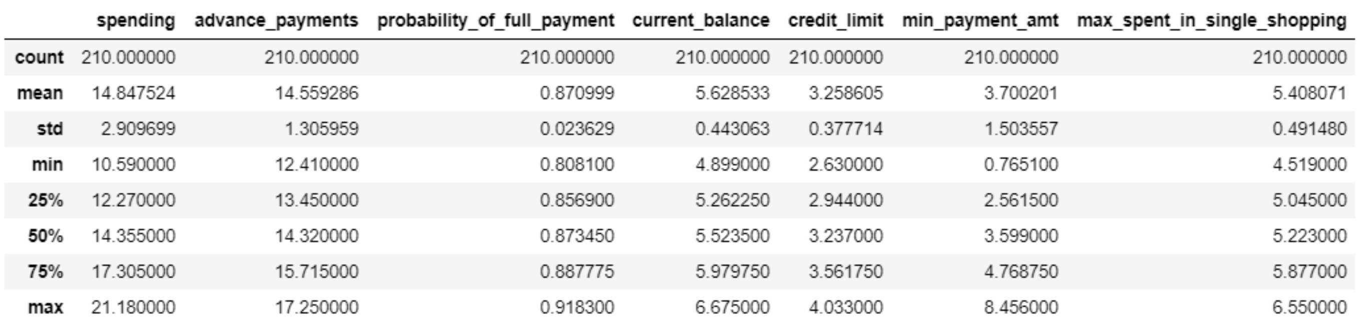
Univariate Analysis:

Plotting Histogram(**Figure: 5)**, Boxplot(**Figure: 6 & 7)** & finding Descriptive summary **(Figure: 4)** for all 7 numerical columns of dataset, as all columns to be plotted are numerical and continuous in nature.

Insights:

- Spending, advance_payments, current_balance & max_spent_in_single_shopping are the columns having **l eft-skewed data.** Other variables are following close to normal distribution.

- There are few outliers/extreme values present for columns: ‘**min_payment_amt**’ & ‘**probability_of_full_payment**’
- The data for variables are not on same scale.‘**Spending’** column has maximum range of 10.59,whereas ran ge of ‘**Probability_of_full_payment**’ is 0.11



	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
count	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000
mean	14.847524	14.559286	0.870999	5.628533	3.258605	3.700201	5.408071
std	2.909699	1.305959	0.023629	0.443063	0.377714	1.503557	0.491480
min	10.590000	12.410000	0.808100	4.899000	2.630000	0.765100	4.519000
25%	12.270000	13.450000	0.856900	5.262250	2.944000	2.561500	5.045000
50%	14.355000	14.320000	0.873450	5.523500	3.237000	3.599000	5.223000
75%	17.305000	15.715000	0.887775	5.979750	3.561750	4.768750	5.877000
max	21.180000	17.250000	0.918300	6.675000	4.033000	8.456000	6.550000

Figure: 4

Plotting histograms for all variables:

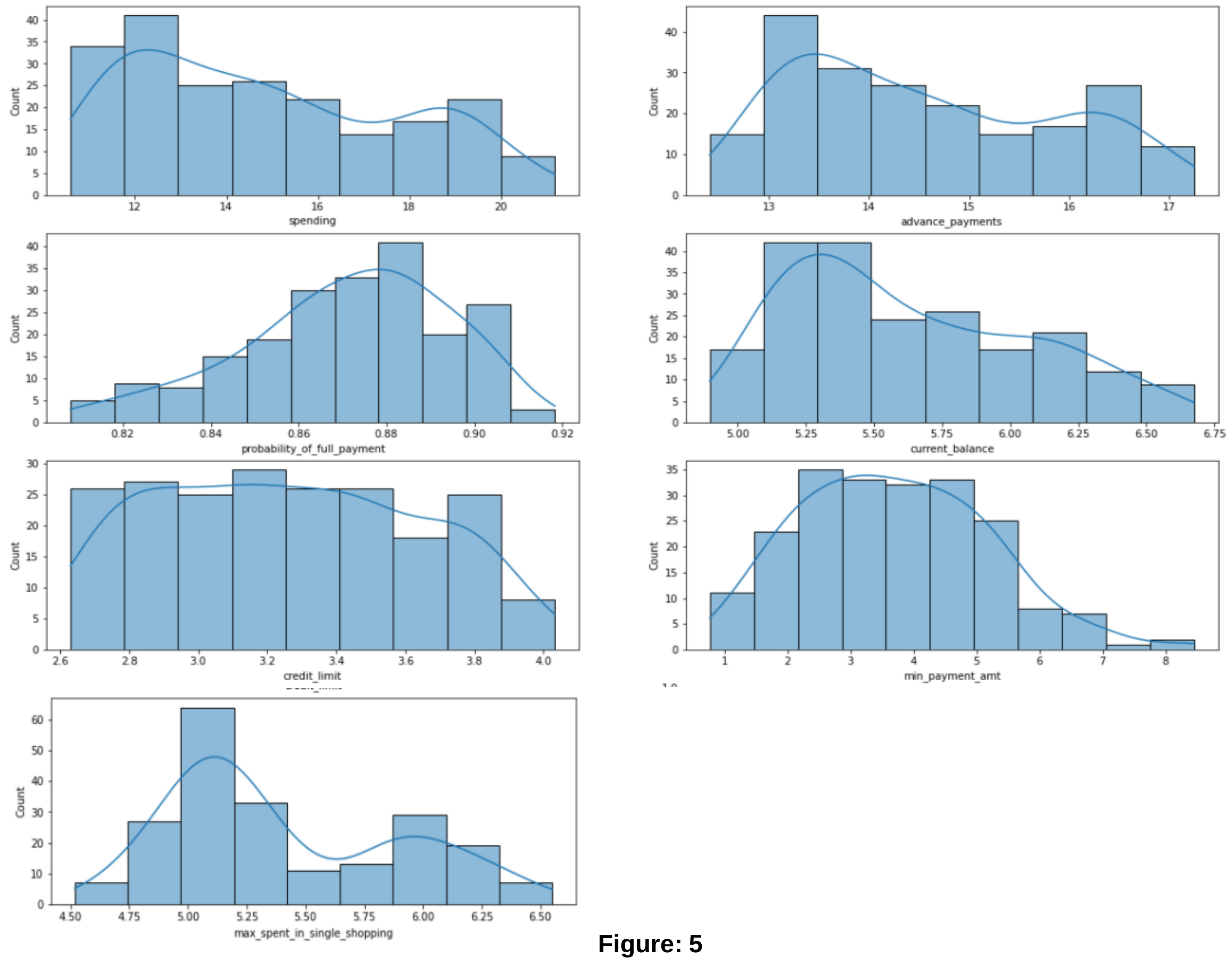


Figure: 5

Plotting Boxplots for all variables:

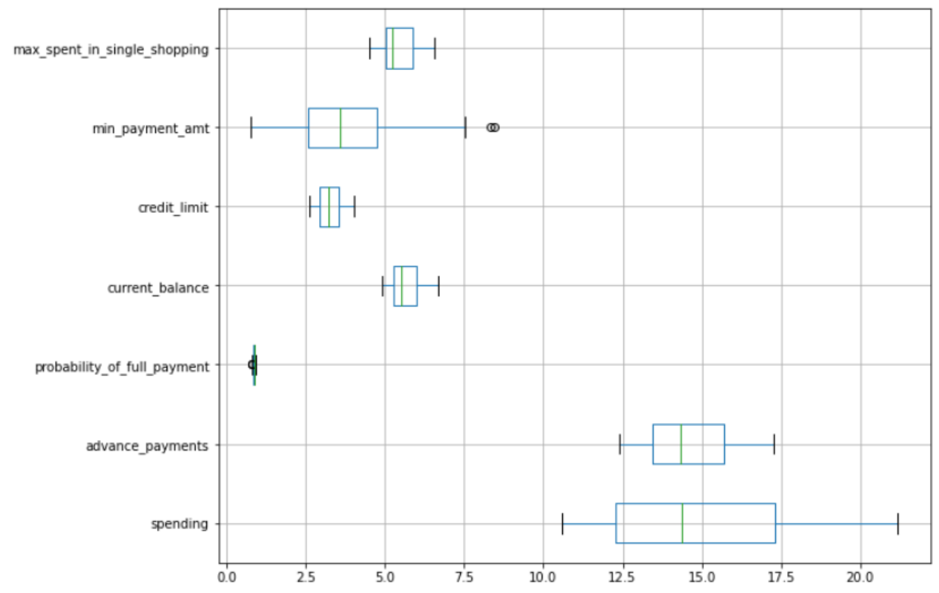


Figure: 6

We have treated outliers as the clustering results are affected by the presence of outliers. This is how the boxplot looks post outlier treatment.

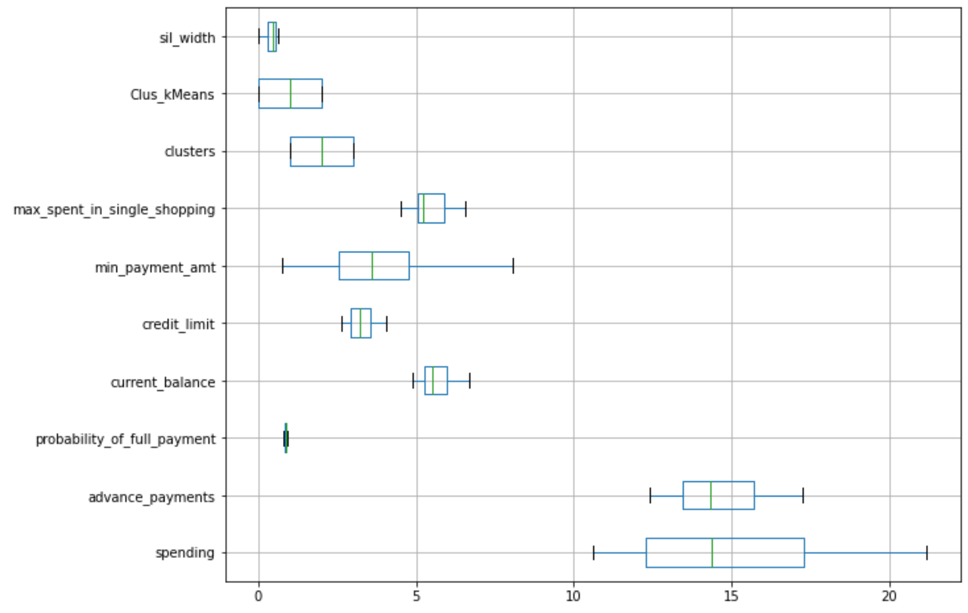


Figure: 7

Bivariate Analysis: Done using Heatmap (**Figure: 8)** & Pairplot (**Figure: 9)** for better readability of all the 7 numerical columns of dataset.

Insights:

Columns '**spending'** & '**advance_payments**' have the strongest correlation = 0.99
Columns '**current_balance'** & '**min_payment_amt**' have the weakest correlation = -0.17
We will be dropping the columns: **'advance_payments' (correlation =0.99)**, **'credit _limit' (correlation =0.97)** from the dataset as these have highest correlation with column: '**spending'**, dropping them won't impact clustering analysis.

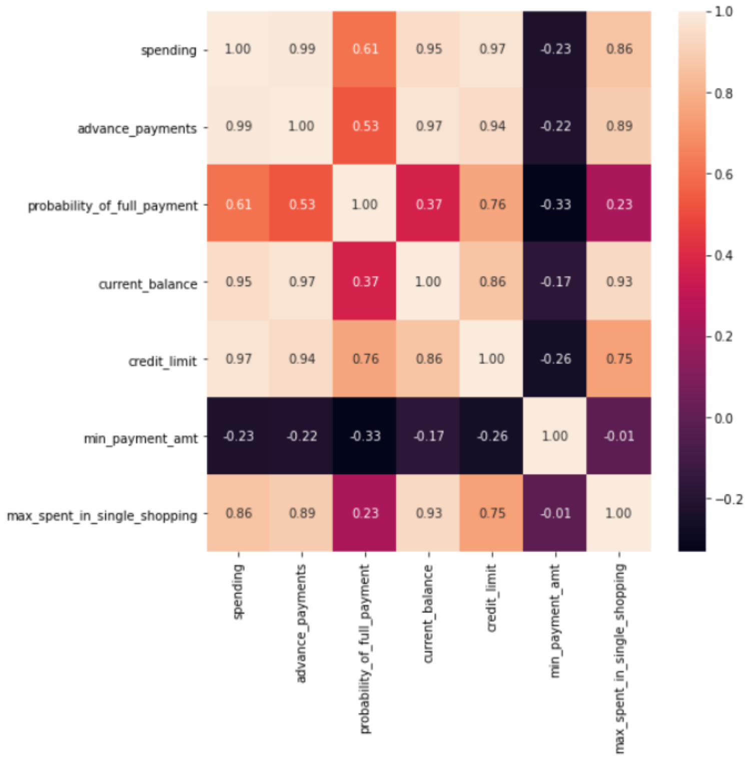


Figure: 8

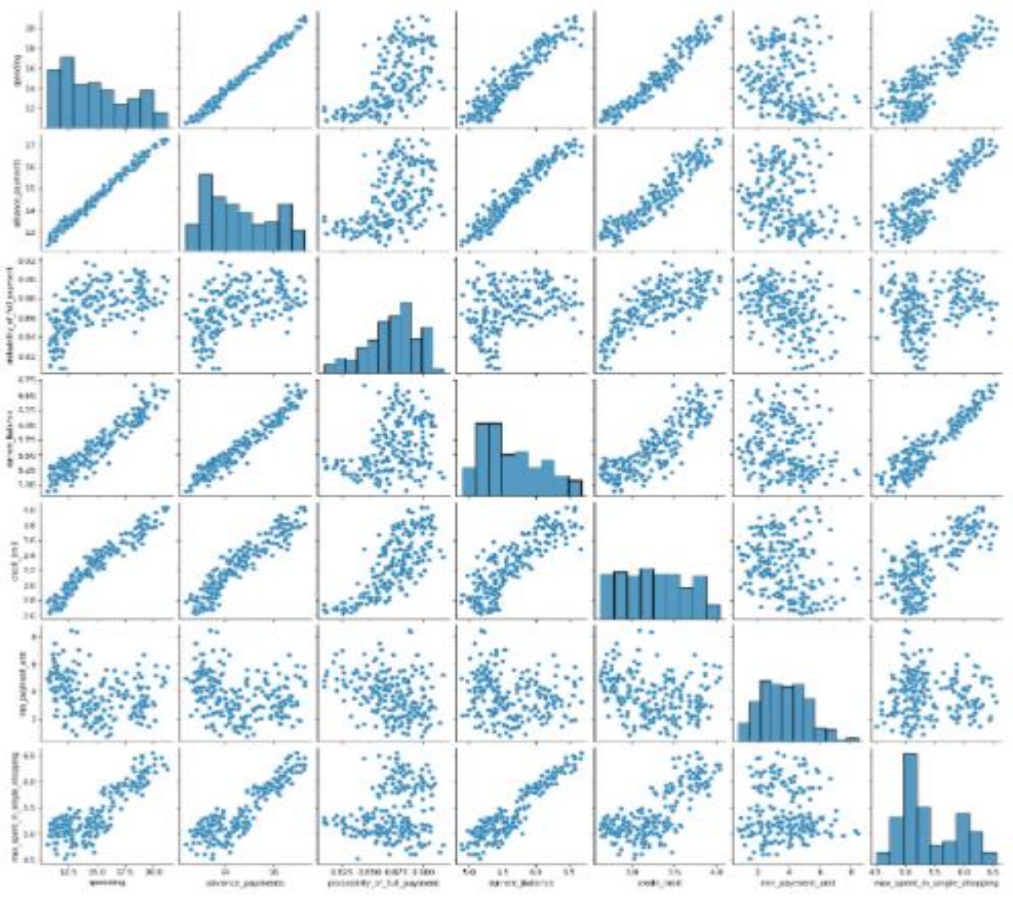


Figure: 9

1.2 Do you think scaling is necessary for clustering in this case? Justify

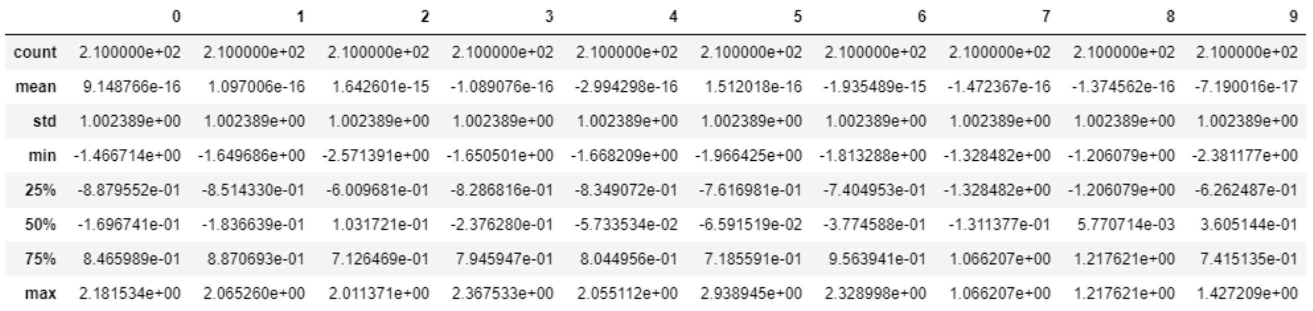
- Since clustering works on distance- based algorithms, here scaling is necessary as the features have different ranges in the data. For clustering, Inter-cluster & Intra-cluster distances are determined between a given pair of observations combining all attributes.

- If scaling is not done, the features with high variance will put more weight on the others with lower variances. Therefore, the cluster formed will be separated towards the variable having higher variance. More is the overall variance, lesser will be the rigidity or cohesiveness of the cluster thus formed.

- In this case, as shown in **Figure: 3,** bank dataset contains features highly varying in range. (For example, range of '**Spending'** = 10.59, whereas range of '**Probability_of_full_payment**' = 0.11).

- The **Standard Scaler method** used from **Sklearn library** is a mean-based scaling method, applied on a near-to-normal data, having less or no outliers, it adjusts the mean as 0 & standard deviation as 1.

This is how our dataset's descriptive summary looks like, post standardization (**Figure: 10)**. The range for different columns are almost same in this case. We would have performed normalization also after this, if the ranges would have differed by large numbers.



	0	1	2	3	4	5	6	7	8	9
count	2.100000e+02	2.100000e+02	2.100000e+02	2.100000e+02	2.100000e+02	2.100000e+02	2.100000e+02	2.100000e+02	2.100000e+02	2.100000e+02
mean	9.148766e-16	1.097006e-16	1.642601e-15	-1.089076e-16	-2.994298e-16	1.512018e-16	-1.935489e-15	-1.472367e-16	-1.374562e-16	-7.190016e-17
std	1.002389e+00	1.002389e+00	1.002389e+00	1.002389e+00	1.002389e+00	1.002389e+00	1.002389e+00	1.002389e+00	1.002389e+00	1.002389e+00
min	-1.466714e+00	-1.649686e+00	-2.571391e+00	-1.650501e+00	-1.668209e+00	-1.966425e+00	-1.813288e+00	-1.328482e+00	-1.206079e+00	-2.381177e+00
25%	-8.879552e-01	-8.514330e-01	-6.009681e-01	-8.286816e-01	-8.349072e-01	-7.616981e-01	-7.404953e-01	-1.328482e+00	-1.206079e+00	-6.262487e-01
50%	-1.696741e-01	-1.836639e-01	1.031721e-01	-2.376280e-01	-5.733534e-02	-6.591519e-02	-3.774588e-01	-1.311377e-01	5.770714e-03	3.605144e-01
75%	8.465989e-01	8.870693e-01	7.126469e-01	7.945947e-01	8.044956e-01	7.185591e-01	9.563941e-01	1.066207e+00	1.217621e+00	7.415135e-01
max	2.181534e+00	2.065260e+00	2.011371e+00	2.367533e+00	2.055112e+00	2.938945e+00	2.328998e+00	1.066207e+00	1.217621e+00	1.427209e+00

Figure: 10

1.3 Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them

Hierarchical clustering seeks to build a hierarchy of clusters using 2 ways: Agglomerative (bottom-up approach) & Divisive (top-down approach)

We are using Dendrogram with ward linkage method from SciPy library to perform Agglomerative clustering here. The linkage stores the various distances at which the various clusters are sequentially merged into a single cluster. We have drawn 2 straight lines, 1 & 2 piercing the Links, to determine the optimum number of clusters or Output partitions obtained (**Figure: 11)**.

Case-1 Piercing using Line1, 2 clusters are formed (as vertical lines are cut)
Case-2 Piercing using Line2, 3 clusters are formed (as 3 links have been broken)
We can see from the dendrogram that the 2 clusters formed (**In Case-1**) are having almost same heights from their child cluster as that of the parent cluster, whereas the 3 clusters formed using Line2(**In Case-2**) are having significant height differences(length of linkage) between their child clusters and parent cluster formed.

Therefore, we can take **Case-2** as final consideration, since the difference between the linkage heights is huge enough to consider noise in the data. More is the difference in distance between 2 clusters formed, higher will be the silhouette score. Hence, the clusters formed will be heterogeneous with larger inter-cluster and smaller intra-cluster distances.

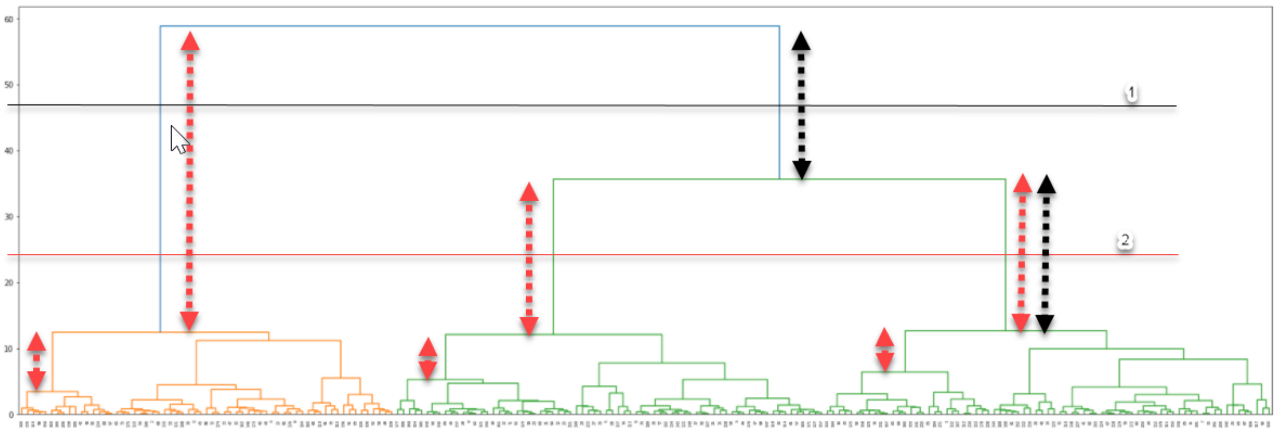


Figure: 11

For better visualization, we have cut the dendrogram formed using ward's method of linkage (**Figure: 12)**, keeping the truncate mode as '**lastp',** which ensures that only last 'p' number of cluster merges (horizontal lines) are shown.

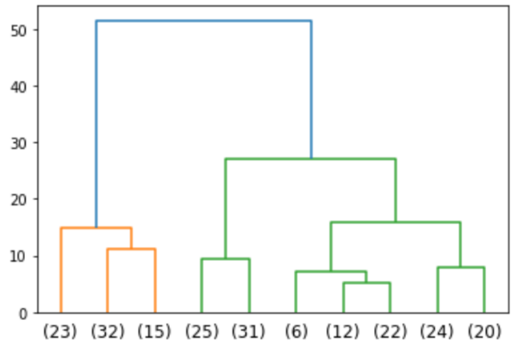


Figure: 12

Next, have also imported **fcluster** function from **scipy** package to form 3 clusters using '**distance'** criterion, which draws a horizontal line at the given distance value to derive the clusters.
This is the array obtained (**Figure: 13)**, which can be attached as a column to the original dataframe for further analysis. (**Figure: 12)**,

```
array([1, 2, 1, 3, 1, 3, 3, 2, 1, 3, 1, 2, 3, 1, 3, 3, 2, 3, 2, 3, 3, 3,
       1, 3, 2, 1, 3, 3, 3, 3, 2, 3, 3, 2, 3, 3, 3, 3, 3, 1, 1, 2, 1, 1,
       3, 3, 3, 1, 1, 1, 3, 1, 1, 1, 1, 1, 3, 3, 3, 1, 2, 3, 3, 1, 2, 1,
       1, 2, 1, 3, 2, 3, 1, 1, 3, 1, 2, 3, 1, 2, 2, 3, 2, 1, 3, 2, 2, 1,
       1, 3, 2, 1, 2, 3, 3, 1, 1, 1, 3, 1, 2, 1, 2, 1, 2, 1, 1, 3, 3, 1,
       2, 2, 1, 3, 3, 1, 2, 3, 3, 1, 3, 3, 3, 3, 2, 2, 1, 3, 2, 2, 3, 2,
       3, 1, 3, 1, 1, 3, 1, 2, 1, 2, 3, 3, 3, 3, 1, 3, 2, 3, 2, 3, 2, 2,
       3, 2, 3, 3, 2, 1, 1, 3, 1, 1, 1, 3, 1, 2, 3, 3, 2, 3, 2, 1, 1, 1,
       2, 3, 2, 3, 2, 3, 3, 2, 1, 1, 3, 2, 3, 3, 3, 2, 3, 1, 2, 1, 1, 3,
       1, 3, 2, 1, 2, 3, 1, 2, 1, 2, 2, 2], dtype=int32)
```

Figure: 13

	spending	probability_of_full_payment	min_payment_amt	max_spent_in_single_shopping	Clus_kMeans	sil_width	Freq
clusters							
1	18.349714	0.884027	3.723400	6.009857	1.000000	0.429482	70
2	14.742500	0.881921	2.575145	5.170571	1.839286	0.386775	56
3	11.999048	0.852926	4.423623	5.064917	0.452381	0.385880	84

Figure: 14

Here is the segmentation of clusters formed:

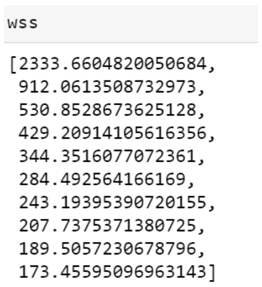
Cluster 0: (Size: 70**) High spending customer group** => This indicates a medium sized group of customers having higher average ‘spending’, higher average ‘probability of full payment’ and high average value of ‘max_spent_in_single_shopping’.

Cluster 1: (Size: 56) **Average spending customer group** => This indicates the smallest size of group of customers having medium average ‘spending’, medium average ‘probability of full payment’ and medium average value of ‘max_spent_in_single_shopping’.

Cluster 2: (Size: 84) **Low spending customer group** => This indicates the largest group of customers having lowest average ‘spending’, lowest average ‘probability of full payment’ and lowest average value of ‘max_spent_in_single_shopping’.

1.4 **Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score. Explain the results properly. Interpret and write inferences on the finalized clusters.**

K-means algorithm tries to partition the dataset into K pre-defined distinct non-overlapping subgroups, where each datapoint belongs to only one group.
In order to find the optimum number of clusters, the elbow method will be used. Here, cluster inertia(**wss**) for a number of clusters between 1 and 10 will be calculated **(Figure: 15).** The rule is to choose the number of clusters where you see a kink or "an elbow" in the graph **(Figure: 16).**



```
wss

[2333.6604820050684,
 912.0613508732973,
 530.8528673625128,
 429.20914105616356,
 344.3516077072361,
 284.492564166169,
 243.19395390720155,
 207.7375371380725,
 189.5057230678796,
 173.45595096963143]
```

Figure: 15

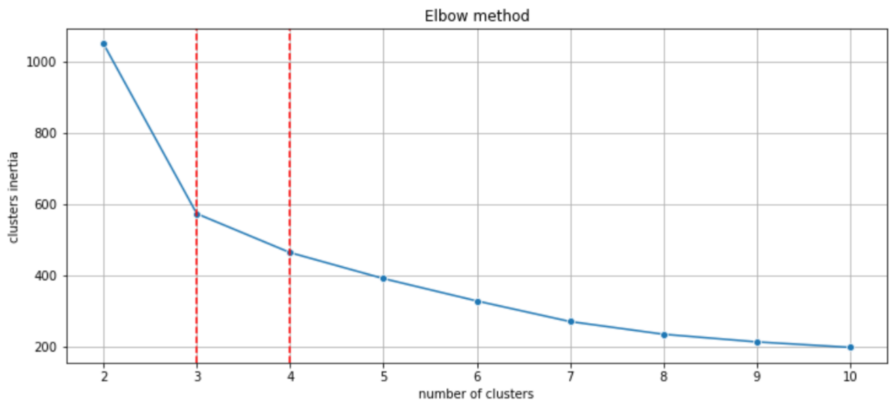


Figure: 16

There is no clear "elbow" visible. A choice of 3 or 4 clusters seems to be fair. Let's see the silhouette scores(**Figure: 17)**

```
silhouette average : 0.3653243409790387


silhouette scores :
 [0.48409159522075446, 0.4936852863569101, 0.43803385184982685, 0.4007709100108336, 0.4108125718523187, 0.38259413447427104, 0.
3776261668238345, 0.367893510472768, 0.3653243409790387]
```

Plotting these using lineplot, we get:

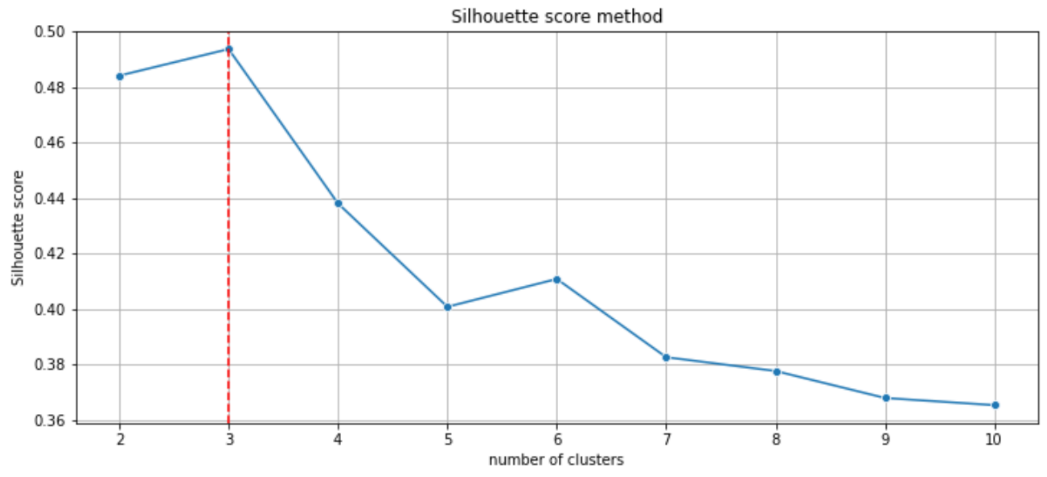
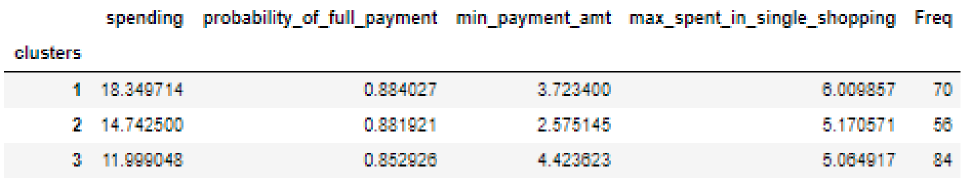


Figure: 17

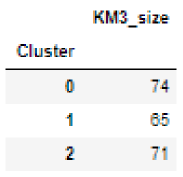
The highest silhouette score is for k=3, so we **assign 3 to the k-means model.**

1.5 Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters.

Below table shows the cluster centroids for the 3 clusters obtained using K-means, for 4 variables, along with the cluster size **(KM3_size:** No of customers falling under that category):



clusters	spending	probability_of_full_payment	min_payment_amt	max_spent_in_single_shopping	Freq
1	18.349714	0.884027	3.723400	6.009857	70
2	14.742500	0.881921	2.575145	5.170571	56
3	11.999048	0.852926	4.423623	5.064917	84



Cluster	KM3_size
0	74
1	65
2	71

Here is the cluster visualization, using scatter plot **(Figure: 18), showing the clusters formed:**

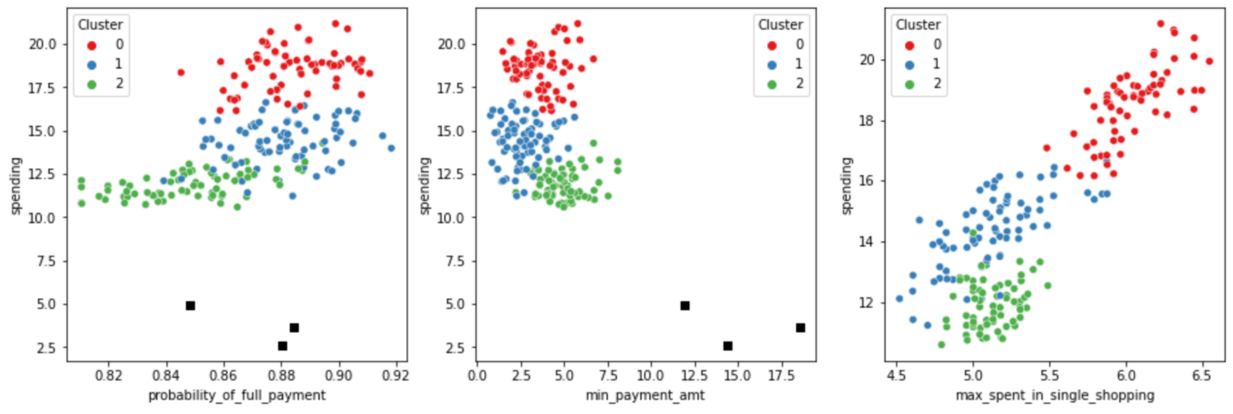


Figure: 18

The 3 clusters can be classified as:

Here is the segmentation of clusters formed:
Max Profitable Group:

- This group consists of the customers who are highly profitable for the business. As the customers have higher average spending, which generates higher revenue.
- Advance payment is highly correlated with spending. Higher spending would lead to higher advance payment, which is again profitable for the company.
- This group also has the highest average value of maximum spent in single shopping, which shows that their shopping amount is the highest, which leads to higher commission for the company.
- For this group, the probability of customer making full payment is highest, which shows that most of them do not rely on loan. They have a sound financial profile.

Average Profitable Group:

- This group has medium spending range, which generates medium revenue.
- This group doesn’t spend a lot in shoppings, which shows that their shopping amount is average, which generates considerable commission for the company.
- Average spending would lead to average advance payment, which generates profit for the company.

Least Profitable Group:

- This group consists of the customers who are least profitable for the business. As the customers have lowest value of average spending, which leads to less revenue.
- Lower spending would lead to lower advance payment, which has negligible impact on the business. This group also has the lowest average value of maximum spent in single shopping, which shows that they spend very less amount on shopping via Credit card.
- For this group, the probability of customer making full payment is lowest, which shows that most of them convert their due amount to EMIs. Converting to EMI generates revenue for the company, but it also holds a considerable risk for the Credit card company.

Recommendations:

- For the profitable customer groups, the company needs to target them to increase their spending amounts, by providing them customized offerings.
- These offerings can be decided after further classification of the profitable groups based on age or their spending profiles. For example, the customers falling under higher age bracket may spend higher on medicals, travel services, etc., whereas the customers falling under lower age bracket may spend more on groceries.
- The bank should encourage the least profitable group customers to use credit card more, by providing them low interest EMIs or additional cashbacks occasionally.